Strategic Re-weighting of U-Net Components in Diffusion Models for Enhanced Speech Enhancement without Retraining

Yuehai Zhang*, Yang Li*, Yuehao Zhao*, and Shoji Makino*
* Waseda University, Japan
E-mail: zhangyuehai@akane.waseda.jp, itsluy@akane.waseda.jp, zhaoyuehao@akane.waseda.jp, s.makino@ieee.org

***Abstract*—Diffusion-based generative speech enhancement (SE) leveraging U-Net architectures has garnered significant attention due to its outstanding performance. In this paper, we explore the inherent potential of U-Net within diffusion models for SE tasks, which can substantially improve generation quality during the diffusion process. Investigating the distinct roles of U-Net components in the SE denoising process, we identify that the main backbone primarily contributes to denoising, while skip connections predominantly introduce high-frequency features into the decoder. This latter effect can cause the network to de-emphasize crucial semantic information from the backbone. Building on this insight, we introduce this method that requires no retraining or fine-tuning. By strategically re-weighting the contributions from U-Net's skip connections and backbone feature maps, our approach seamlessly integrates into existing diffusion models and demonstrably enhances speech enhancement performance with minimal code modification.**

I. Introduction

Speech enhancement (SE) aims to recover clean speech from audio recordings degraded by noise [1]. It serves as a crucial component in various speech processing systems, including speech recognition and semantic communication [2]. Recently, generative SE approaches have garnered significant attention due to their strong denoising capabilities. In this paradigm, generative models learn the intrinsic properties of clean speech, such as its underlying spectral and temporal structures. This enables them to reconstruct speech signals by focusing on these learned characteristics, thereby aiming to separate them from interfering noise factors.

Among prominent generative SE models, diffusion models have recently attracted significant attention, largely due to their notable successes in wide domain [3]–[10]. Inspired by non-equilibrium thermodynamics, diffusion models operate by progressively transforming data into noise during a forward process, while a neural network learns to reverse this noise-adding procedure. Conditional diffusion models [9], for instance, utilize noisy spectrograms as input conditions. However, their objective function often assumes that the global distribution of additive noise conforms to a standard white Gaussian distribution, an assumption that may not align with real-world noise statistics. In contrast, score-based diffusion models, grounded in stochastic differential equations (SDEs), facilitate a fully generative training process without imposing prior assumptions on the noise distribution. A common challenge with diffusion models is the computationally intensive nature of their reverse diffusion process (decoding). To mitigate this and reduce the number of reverse diffusion steps, several studies have proposed combining a predictive model with the generative framework [8], [11].

The neural network central to the reverse diffusion process, tasked with iteratively denoising the signal, is commonly implemented using a U-Net architecture [12]. This architecture is favored due to its encoder-decoder structure, which extracts hierarchical contextual information, and its skip connections, which facilitate the propagation of fine-grained, high-frequency details from the encoder to the decoder. In the context of diffusion models for speech enhancement, the U-Net effectively estimates the noise component at each step of the reverse diffusion, progressively refining the output towards the clean speech target. The adoption of U-Net has been instrumental in the success of diffusion models for tasks requiring detailed signal reconstruction, demonstrating strong performance in generating high-fidelity speech.

However, despite its empirical success, the precise mechanisms through which the U-Net's components contribute to speech enhancement remain largely unexplored. It is not yet well understood how its backbone features and skip-connection features individually and collectively function to separate speech from noise. Inspiration for a deeper analysis comes from the text-to-image generation domain, where a recent study, FreeU [13], delved into the roles of the U-Net architecture. It revealed that strategically modulating the contributions of different feature levels can significantly enhance the quality of generated images by dynamically adjusting the balance between low and high frequency components.

Motivated by this insight, we systematically investigate the low and high frequency components changes in the context of speech enhancement. Building on our findings in Fig. 1, we propose a weighting method to improve the SGMSE+ model [7]. This method adjusts the influence of the backbone and skip-connection features during the inference stage, delivering a "plug-and-play" performance boost without requiring any model retraining or fine-tuning. Crucially, this approach is designed to be generalizable, holding the promise of applicability to a wide range of other U-Net-based generative models

979-8-3315-7206-8/25/$31.00

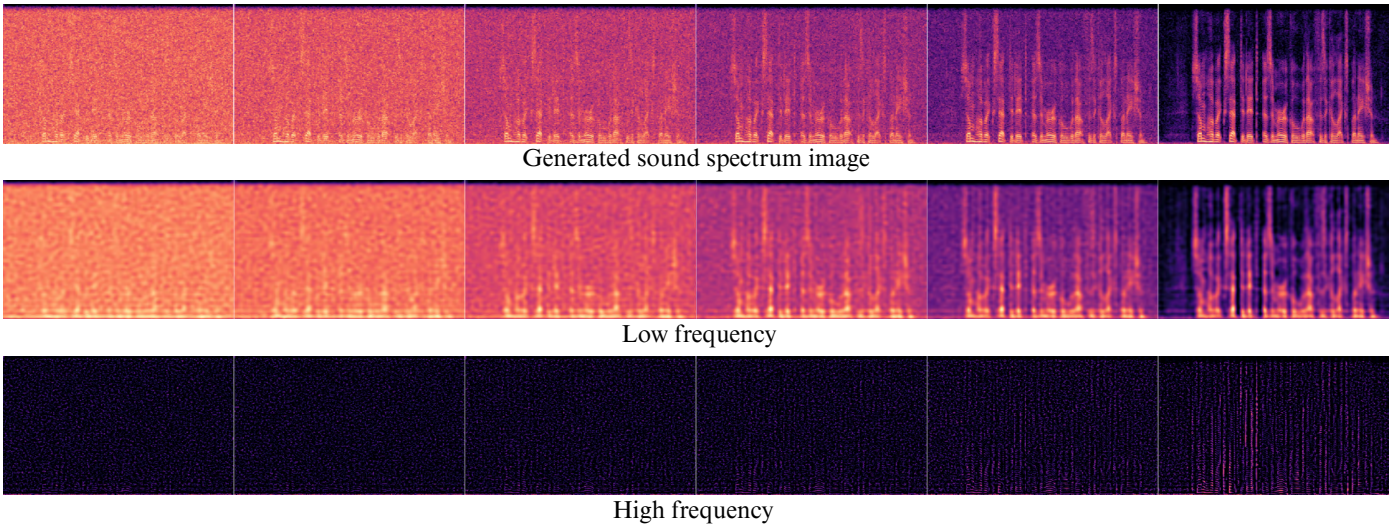


Fig. 1. The denoising process. The top row depicts the progressive refinement of the spectrogram over 30 iterations. The second and third rows display the corresponding low and high frequency components of the signal, respectively. These time-domain components are generated at six-iteration intervals by applying an Inverse Short-Time Fourier Transform (ISTFT) to the spectrogram. It is evident that the majority of the noise is concentrated in the low-frequency band, which, as a result, undergoes more significant changes during the denoising procedure.

for speech enhancement.

II. Diffusion Model with U-Net architecture

A. Diffusion model

To achieve the SE task, forward and reverse processes are defined [14]. Audio processing is performed in the frequency domain, where the real and imaginary parts of the complex signal are treated as two separate channels. In the forward process, the clean spectrogram x_0 is gradually transformed into the noisy speech representation x_T. The variable $\mathrm{t} \in [0, T]$ represents a continuous time step for the process. In the reverse process, the noisy signal x_T is restored to x_0 using a score, s_θ, generated by a score model. The forward process is a stochastic diffusion process, $\{\mathrm{x_t}\}_{\mathrm{t}=0}^{T}$, modeled as the solution to a linear Stochastic Differential Equation (SDE) [7]:

$$\mathrm{dx_t} = \gamma(\mathrm{y} - \mathrm{x_t})\mathrm{dt} + g(\mathrm{t})\mathrm{dw}, \tag{1}$$

where y is the noisy speech, γ is a constant controlling the drift towards y, w is the Wiener process, and $g(\mathrm{t})$ is the diffusion coefficient that controls the added noise:

$$g(\mathrm{t}) := \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\mathrm{t}}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}. \tag{2}$$

Here, $\sigma_{\min}$ and $\sigma_{\max}$ define the noise schedule. The reverse process is also an SDE, which restores the clean signal:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}. \tag{3}$$

The significant component is the score, $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{y})$, which is a vector field representing the direction of steepest ascent in the data's log-likelihood. This score is estimated by a score model $s_\theta(\mathrm{x_t}, \mathrm{y}, \mathrm{t})$ trained with an MSE loss.

B. U-net

The structure of the score model $s_{\boldsymbol{\theta}}(\mathbf{x}_\mathrm{t}, \mathbf{y}, \mathrm{t})$ used for SGMSE+ [7], the baseline model, is called a Noise Conditional Score Network (NCSN++). It receives $\mathbf{x}_\mathrm{t}, \mathbf{y} \in C^{\mathrm{F}\times\mathrm{N}}$, and t as inputs and estimates a score as output. To deal with complex signals, SGMSE+ handles their real and imaginary parts as separate real signals.

In NCSN++, the score is primarily estimated by a U-Net block [12], which comprises an encoder, a bottleneck, and a decoder. The encoder processes feature maps from the input signals $(\mathbf{x}_\mathrm{t}, \mathbf{y})$, progressively reducing their spatial dimensions. To stabilize training, each encoder layer is also conditioned on down-sampled versions of the input signals provided by a Progressive Down-Sampling (ProgDown) block. Conversely, the decoder gradually up-samples the feature maps to their original size. Its intermediate outputs are sent to a Progressive Up-Sampling (ProgUp) block, which combines them to produce the final score estimate. Both ProgDown and ProgUp utilize fixed, non-trainable functions. As the largest component in NCSN++, the U-Net accounts for the majority of the computational cost during training and inference [15].

C. Denoising ability of diffusion U-Net

Inspired by previous research, FreeU [13], on text-to-image tasks, which explored the distinct roles of U-Net components. FreeU suggested that the U-Net backbone acts as a denoiser, while skip connections preserve high-frequency details. In Fig. 1, we investigate the U-Net's denoising mechanism in the audio domain; an experiment was conducted to visualize the progressive refinement of a spectrogram. The analysis reveals that the denoising operation is primarily concentrated in the low-frequency band. The low-frequency component undergoes a significant transformation, evolving from a noise-saturated

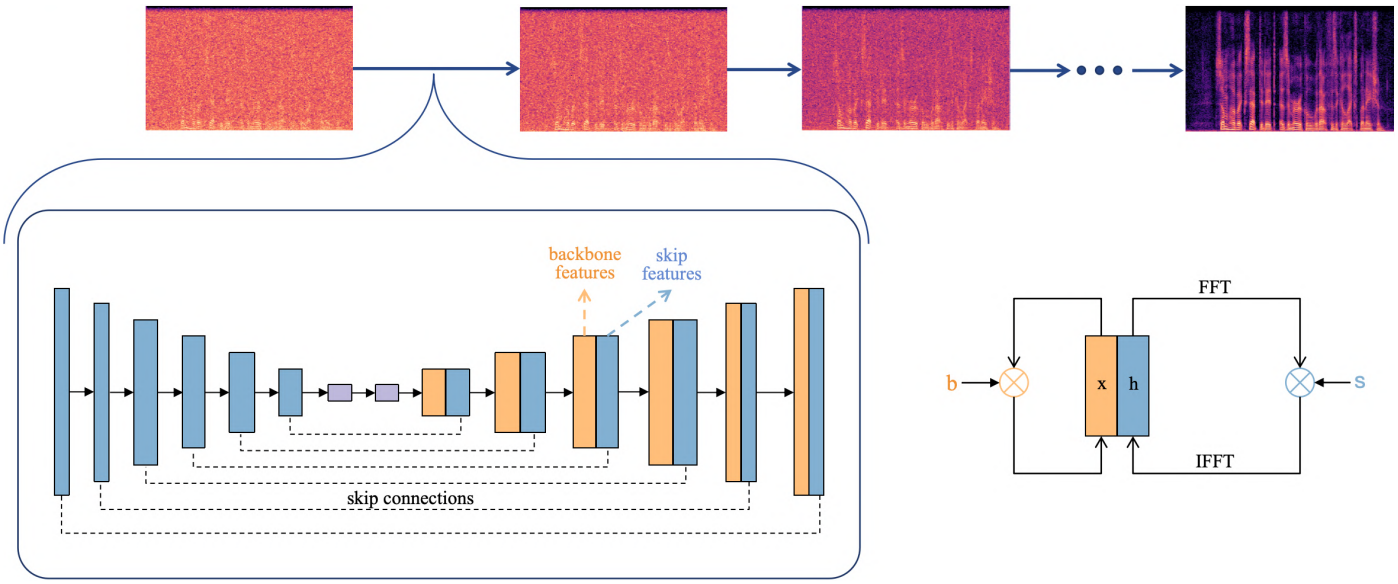


Fig. 2. **Framework of the proposed method.** The proposed approach utilizes a U-Net architecture, which is enhanced at a decoder stage. Specifically, the backbone feature map x is scaled by a factor b, while the corresponding skip-connection feature h is modulated in the frequency domain: it is transformed via FFT, scaled by a factor s, and then returned to the time domain via IFFT.

state to a structured signal representation. In contrast, the high-frequency component exhibits minimal change, remaining sparse throughout the process. This observation indicates that the majority of the initial noise resides in the lower frequencies, and the model effectively targets this band for denoising.

Based on this finding, we introduce a method of strategically re-weighting U-Net components by applying scaling factors b and s to the feature maps of the backbone and skip connections.

III. Proposed Methods

Inspired by FreeU [13], we introduce a simple and effective method to improve the generation quality of U-Net based models without additional training or fine-tuning by modulating features during inference. The U-Net's denoising performance is augmented through a dual modulation of its features. The backbone feature maps (x_l) are strengthened by applying a scaling factor b_i to the first half of the channels. Concurrently, the skip connection features (h_l) are attenuated. The modification to the backbone is defined as:

$$\mathrm{x}'_{l,i} = \begin{cases} \mathrm{x}_{l,i} \odot b_i, & \text{if } i < \mathrm{C}/2, \\ \mathrm{x}_{l,i}, & \text{otherwise.} \end{cases} \tag{4}$$

A spectral modulation technique is concurrently applied to the skip features h_l to mitigate the loss of high-frequency details. The core of this technique is the attenuation of low-frequency components in the Fourier domain. The procedure involves transforming the feature, applying a frequency mask $\beta_{l,i}$, and then performing an inverse transform:

$$\mathrm{h}'_{l,i} = \mathrm{IFFT}(\mathrm{FFT}(\mathrm{h}_{l,i}) \odot \beta_{l,i}). \tag{5}$$

The mask $\beta_{l,i}$ applies a scaling factor s to frequencies below a radius threshold r_{thresh}, The radius threshold r_{thresh} is a hyperparameter that defines the boundary for which frequency components are scaled. And different values would alter the range of frequencies being adjusted. Following setup of FreeU, we set $r_{\text{thresh}} = 1$ in our experiments:

$$\beta_{l,i}(\mathrm{r}) = \begin{cases} s_l & \text{if } r < r_{\text{thresh}}, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

Finally, the modified features x'_l and h'_l are concatenated for subsequent layers. It's a practical, computationally inexpensive solution that can be seamlessly integrated into existing diffusion models to enhance performance.

IV. Experiments

A. Datasets

The experiments are conducted on the Voice-Bank+DEMAND dataset, a widely recognized benchmark for speech enhancement tasks. The clean speech signals are sourced from the VoiceBank corpus [16], which comprises recordings from 30 distinct speakers. The noise signals are drawn from the DEMAND database [17], containing a diverse set of real-world, non-stationary noises.

To ensure a fair and direct comparison, we strictly adhere to the experimental setup established by the SGMSE+ baseline. The training set is synthesized by mixing utterances from 28 speakers with a variety of noises at multiple Signal-to-Noise Ratios (SNRs). The test set consists of mixtures created from the remaining two speakers, who were unseen during the training phase. For our qualitative analysis, audio samples were randomly selected from this standard test set

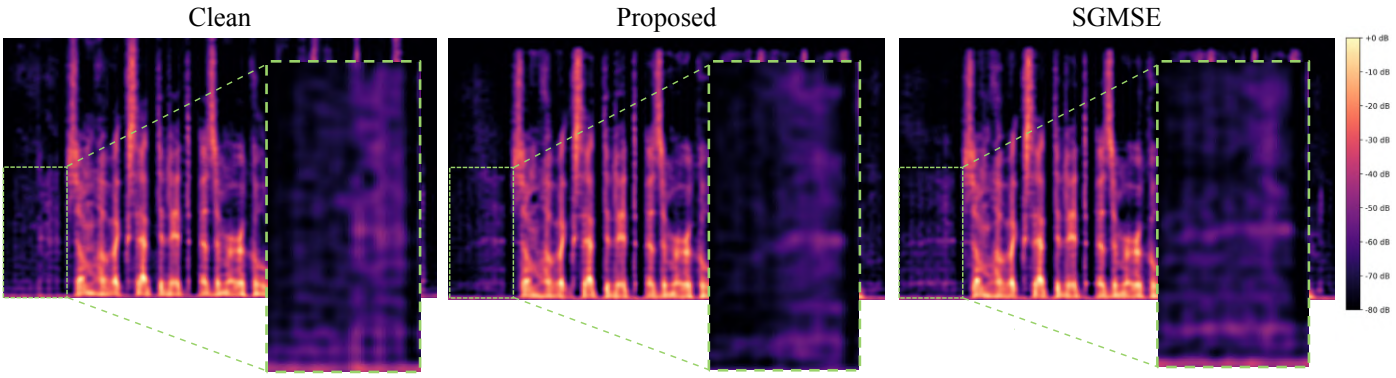


Fig. 3. Spectrogram comparison.

to facilitate a representative and unbiased evaluation of the model's performance.

B. Evaluation Metrics

To comprehensively assess the quality of the generated enhanced speech, we employ a suite of standard objective metrics. Specifically, we use Perceptual Evaluation of Speech Quality (PESQ) [18] to measure perceptual quality, Extended Short-Time Objective Intelligibility (ESTOI) [19] to predict speech intelligibility, and Signal-to-Artifact Ratio (SI-SAR) [20] to quantify signal fidelity. To ensure experimental reliability and a fair comparison, all metrics were computed using the identical methodology as the SGMSE+.

C. Results

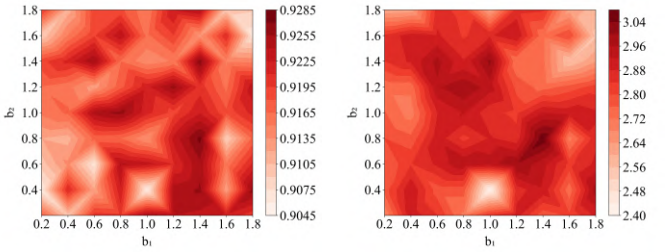
We selected the audio sample `p232_013` from the VoiceBank dataset and applied our proposed methods for enhancement. A comparison with the SGMSE+ baseline is presented in Table I. The "Proposed 1" and "Proposed 2" variants correspond to different parameter settings for the variables $s$and b. Notably, our proposed method achieves superior speech enhancement results without incurring additional computational costs or requiring retraining.

TABLE I
OBJECTIVE EVALUATION RESULTS.

Method	PESQ	ESTOI (%)	SI-SAR (dB)
SGMSE+	2.85	92.15	15.90
Proposed 1	3.05	92.14	15.68
Proposed 2	**3.25**	**93.39**	**16.43**

The comparison in Fig. 3 demonstrates that our proposed method preserves more details of the clean signal than the baseline.

To assess the robustness and applicability of the proposed method, we constructed an evaluation set by randomly sampling 80 clips from the Voicebank corpus. The proposed algorithm was then applied to this set, with its parameters for each clip. A comparative analysis against the SGMSE+ baseline was conducted, and the results are summarized in Table II. As shown, our method achieves a PESQ score of 3.31 and an ESTOI of 88.25%, surpassing the baseline in both metrics while maintaining a comparable SI-SAR score. These improvements underscore the efficacy of our approach.



(a) ESTOI (b) PESQ

Fig. 4. Ablation study on backbone scaling factor b.

TABLE II
EVALUATION RESULTS ON A SUBSET OF THE VOICEBANK DATASET.

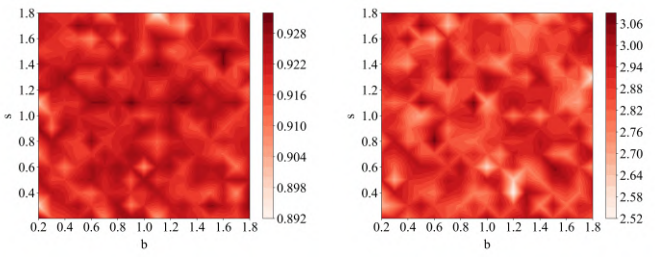
Method	PESQ	ESTOI (%)	SI-SAR (dB)
SGMSE+	3.02	87.84	18.44
Proposed	**3.31**	**88.25**	18.44

D. Ablation Studies

For the ablation study, we deeply investigate the influence of the parameters b_i on the denoising ability. We performed an ablation study to scrutinize the impact of the backbone scaling factors, b_1 and b_2.

The results, presented in Fig. 4, demonstrate that enhancement performance is highly sensitive to these parameters. A key finding is that the optimal configurations lie in off-diagonal regions, indicating that an asymmetric scaling strategy is most effective. Conversely, simultaneously increasing both factors proves to be a suboptimal strategy, confirming that a naive, uniform amplification of backbone features fails to unlock the model's full potential.

Then We evaluated a global application of the feature scaling method from FreeU [13] across all layers. This experiment



(a) ESTOI (b) PESQ

Fig. 5. Ablation study on all layers with parameters b and s.

produced a highly unstable performance, as shown in Fig. 5, with extreme sensitivity to the scaling factors, making hyperparameter tuning infeasible. Based on this finding that a global application is suboptimal, we restricted the scaling to a single layer (layer 2). This selective approach resulted in more robust performance enhancement, as detailed in Table II.

V. Conclusions

In this paper, we introduce an inference-only weighting method to enhance the performance of U-Net architectures in diffusion-based speech enhancement. By strategically rebalancing the distinct contributions of the U-Net's semantic backbone and its high-frequency skip connections, our approach achieves a significant improvement in speech quality. This method functions as a zero-cost, plug-and-play module for pre-trained models, eliminating the need for retraining or fine-tuning. Promising avenues for future investigation include extending this principle to other U-Net-based generative audio tasks and developing more adaptive weighting strategies.

References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*. Cham, Switzerland: Springer, Jul. 2013.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer, 2006.

[3] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proc. Interspeech*, 2024, pp. 1175–1179.

[4] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open J. Signal Process.*, 2024.

[5] Y. Li, Y. Sun, and P. P. Angelov, "Complex-cycle-consistent diffusion model for monaural speech enhancement," in *Proc. AAAI.*, vol. 39, 2025, pp. 18 557–18 565.

[6] W. Tai, Y. Lei, F. Zhou, G. Trajcevski, and T. Zhong, "Dose: Diffusion dropout with adaptive prior for speech enhancement," in *NeurIPS*, vol. 36, 2023, pp. 40 272–40 293.

[7] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[8] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2724–2737, 2023.

[9] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE ICASSP*, 2022, pp. 7402–7406.

[10] Z. Guo, J. Du, C.-H. Lee, Y. Gao, and W. Zhang, "Variance-preserving-based interpolation diffusion models for speech enhancement," pp. 1065–1069, 2023.

[11] H. Shi, K. Shimada, M. Hirano, *et al.*, "Diffusion-based speech enhancement with joint generative and predictive decoders," in *Proc. IEEE ICASSP*, 2024, pp. 12 951–12 955.

[12] P. O.Ronneberger and T.Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[13] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "Freeu: Free lunch in diffusion u-net," in *Proc. CVPR*, 2024, pp. 4733–4743.

[14] Y. Li, Z. Qiu, and S. Makino, "Global convolutional-block-attention-module-based diffusion model for speech enhancement," in *Proc. NCSP*, Pulau Pinang, Malaysia, Feb. 2025.

[15] R. Kimura, T. Nakatani, N. Kamo, *et al.*, "Diffusion model-based mimo speech denoising and dereverberation," in *Proc. IEEE ICASSP. Workshop*, 2024, pp. 455–459.

[16] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. ISCA Speech Synth. Workshop*, 2016, pp. 146–152.

[17] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multichannel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, Salt Lake City, UT, USA, 2001, pp. 749–752.

[19] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *Proc. IEEE ICASSP*, Brighton, UK, 2019, pp. 626–630.